DOCUMENT RESUME

ED 454 855 IR 058 146

AUTHOR	Lynch, Clifford
TITLE	The New Context for Bibliographic Control in the New Millennium.
PUB DATE	2000-11-00
NOTE	11p.; In: Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web (Washington, DC, November 15-17, 2000); see IR 058 144.
AVAILABLE FROM	For full text: http://lcweb.loc.gov/catdir/bibcontrol/lynch_paper.html.
PUB TYPE	Opinion Papers (120) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Access to Information; *Cataloging; Information Networks; *Information Retrieval; *Information Seeking; Nonprint Media; Online Systems; Relevance (Information Retrieval); Standards
IDENTIFIERS	*Digital Data

ABSTRACT

This paper considers the ways in which information finding is changing in a world of digital information and associated search systems, with particular focus on methods of locating information that are distinct from, but complementary to, established practices of bibliographic description. The following three general approaches to identifying potentially relevant information are described: through bibliographic surrogates that represent an intellectual analysis and description of aspects and attributes of a work; through computational, content-based techniques that compare queries to parts of the actual works; and through social processes that exploit the opinions and actions of communities that author, read, and evaluate works, as well as the information seeker's view of those communities. Ways that computational content-based retrieval can help information seekers and techniques for making non-textual materials available are discussed. Three areas are explored as part of the context for the new bibliographic control: (1) bibliographic control is not just about rules and practices, it also depends upon a complex infrastructure of authority files and classification structures; (2) the networked information environment has a democratizing and empowering character; and (3) as part of the massive migration of content to digital form, we are approaching a crucial point in standards-setting. (MES)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Comments by Jennifer Trant

The New Context for Bibliographic Control In the New Millennium

ED 454 855

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

B. Wiggins

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Clifford Lynch

Final version

1

This text is based on a dinner speech given in the Great Hall of the Jefferson Building of the Library of Congress on the evening of November 16, 2000 as part of the Bibliographic Control for the New Millennium conference.

The broad reading public, from scholars to students, researchers to recreational readers, wants and needs to find works of relevance to their interests. Enabling the identification of such works meeting these needs is not the only purpose of bibliographic control, but it is certainly one of the most important and most widely relied-upon. It is the most visible, and, to a great extent, the reason why there is support for the very substantial ongoing investment in bibliographic control. But the practices of information finding are changing in a world of digital information and computer-based search systems. Within the library community we have placed great emphasis on the impact of on-line public access catalogs and abstracting and indexing databases designed to be searched by the general public rather than specially trained intermediaries for several reasons. These systems arrived early; they have been deployed for about two decades on a reasonable scale and are now deployed almost ubiquitously. They are effective and well-received by their users; they represent a very significant improvement in the quality of access to library collections. And the library community is, at some fundamental level, comfortable with these systems; they empowered users of traditional, mostly print library collections and leveraged and reinforced the traditional philosophies and approaches to bibliographic control.

But these systems were not fundamentally revolutionary or transformational; they represent a process of modernization through automation, of measured evolution. The real revolution in access is just starting to arrive; this is going to be driven by the availability of massive amounts of content directly in digital form rather than print, and by the emergence of network-based computer systems that provide an environment not just for identifying content (which historically existed in print form and was used offline, independent of systems like online catalogs) but for its subsequent actual use and analysis within the access system. Indeed, the same computer systems that provide identification, access and an environment for reading and use may also serve as collaborative environments for new authoring . This is the new context for bibliographic control, and we ignore it at our peril; it will certainly reorder priorities for investment in bibliographic control practices and it will change the way that cataloging information, for example, is used and the purposes to which it is put in support of seeking relevant information.

I will focus here on methods of locating information that are distinct from, but complementary to,

IR058146

BEST COPY AVAILABLE

established practices of bibliographic description. A full understanding of these developments is essential in re-thinking bibliographic control in the new millennium, because they fundamentally change the roles and importance of bibliographic metadata in the information discovery processes.

There are three general approaches to identifying potentially relevant information:

- through bibliographic surrogates, that represent an intellectual analysis and description of aspects and attributes of a work; through computational, content-based techniques that compare queries to parts of the actual works themselves (or to computationally-derived surrogates for the works);
- through social processes that exploit the opinions and actions of communities that author, read, and evaluate works, and the information seeker's view of those communities of people involved.

The first approach is familiar, and forms the basis of catalogs and abstracting and indexing, and more recently online catalogs and similar systems. I will return to the question of how this changes in the new digital environment shortly.

The third approach is also familiar, in the form of book and article reviews, and suggestions from colleagues, and more recently citation indexing, but is now seeing a great creative expansion in the digital world, with its ability to create and aggregate world-wide communities of interest and to track the behavior of users within these communities. In this area we find fascinating and exciting current developments such as reccomender systems and collaborative filtering, which sometimes translate tracked behavior into implied ratings and which also permit the development of highly democratic, participatory and distributed explicit rating systems. We can also see here developments in trust and reputation management systems that begin to allow individuals to extend ideas about which opinions they trust and respect from limited and slowly changing circles of friends and colleagues to large dynamic global network-based communities that include many relative strangers. It is interesting to note that while this is a very powerful approach in support of individual information seekers, it is of much less use for intermediaries and for those concerned with the stewardship of collections.

The second approach is fundamentally new and indeed possibly only in the digital world, where techniques based on full text searching form the basis of todayÕs web search engines. The key point to recognize is that within a very few years virtually all new material, and an ever-growing amount of previously published material is going to be available in digital form as a routine matter. We need to recognize that in the new millennium, for digital materials, effective content-based computational techniques will be a very inexpensive, ubiquitous, default means of searching, available virtually the instant that the content is first distributed or published, and that powerful socially based approaches will also be widely available at little cost, as a byproduct of the authoring, dissemination and subsequent use of the works. The information identification support provided by human-based intellectual bibliographic control, which is intrinsically more costly and often available only after some delay following dissemination of a work, will have to compete with these other methods of finding relevant information, and do so with enough success to justify its costs.

It is worth noting some of the controversies surrounding full content searching and also worth recognizing some of its very real limitations.

Starting in the 1950s or thereabouts, a group of computational and information scientists began to develop a wide range of technologies to support effective full text retrieval without the use of bibliographic surrogates; their vision was that this would lead to far more effective and flexible searching and information location capabilities than bibliographic surrogates offered, and that ultimately it would also be far less expensive as the cost of computer cycles and storage continued to decline. Bluntly, if they achieved this vision, there would no longer be much need for bibliographic control, at least in support of information finding -- a very threatening prospect to many in the traditional library community. There were two major problems, however. Developing the technology to a reliable, robust level of maturity turned out to be extraordinarily difficult (as some people from the bibliographic control world enjoyed pointing out from time to time as the latest over-hyped technology developments surfaced). And even if the technology could be made to work, only an miniscule, insignificant proportion of the important literature existed in machine readable form so that the computation technology could be applied to it. Just about everything important was only available in print, while the researchers played with small, specially-constructed test databases.

Fifty years later, and after the investment of billions of dollars and countless years of human effort in research and development, the world has changed a great deal. The vision still hasn't been fully achieved (computers still have a lot of trouble deciding what texts are really "about", in a meaningful way, for example). But there is compelling evidence from full text searching systems (including web search engines) that content-based searching offers some capabilities that are completely unattainable through the use of bibliographic surrogates, and are often very valuable. Imagine being able to find every document that mentions a certain specific person, place or thing (right down to the passage in the document), to take one simple example. This is impossible with bibliographic surrogates (which weren't designed to solve this problem) but for many research needs it is absolutely revolutionary.

Researchers continue, appropriately, to push towards the vision and also to explore new ways that content-based retrieval can help information seekers; my personal view is that it will be a long time before they can replace human intellectual analysis by computation. But it is clear that current content-based systems complement traditional bibliographic control in supporting information seeking and provide capabilities that are not otherwise available. It is time -- indeed past time -- for the bibliographic control community to recognize the legitimacy of computational content-based retrieval and to understand its strengths and its contributions to information access, and also to look with an open mind at types of queries and classes of content where computational methods may compare favorably to bibliographic control based approaches, or may at least be "good enough", particularly given their very low cost.

As to the other objection, the paucity of content in digital form, as already discussed virtually all content is moving to digital form rapidly. The Web isn't' a test database -- it's a real-world collection of an enormous amount of information, some of it of great quality, importance and timeliness. There are some technical issues, and also some messy intellectual property issues (in part technical, in part legal and business) that will need to be resolved in order to make sure that the output of traditional publishing

processes is available for indexing and searching by these computational systems (in the same way that it has been to catalogers, abstracters, indexers and reviewers), and this will take time and probably cause some considerable disruption and uproar along the way -- but this is another set of issues, for discussion another time. The key point is that we have now reached a "critical mass" of digital materials, and this will only grow, and this content will become available for computational indexing and retrieval.

There is one other essential point I must make here. Thusfar, while I have sometimes used the general term "content-based retrieval" what I have mainly been talking about is textual information. One of the great potentials of the digital environment is to elevate images, sound recordings, video, interactive simulations and other types of materials to a much more mainstream role in discourse, communication and the representation and capture of knowledge and of events than they have enjoyed up till now. We are already starting to see this happen; digital articles, term papers, or business communications can incorporate these nontextual components much more casually than their print predecessors. Tremendous amounts of audio and video are being routinely captured as a byproduct of various events and subsequently made available.

The best techniques that we have for making these kinds of non-textual materials available is to use human intellectual analysis to attach words to them (ideally within a structured descriptive or analytic context), and then to use these words as surrogates; much of this is essentially bibliographic analysis and control, or broader scholarly analysis, description and classification. Other techniques for using words to gain leverage on non-textual materials have a more mechanical character; transcribing talks, or creating closed caption tracks for video. There have been tremendous investments in technologies to make content accessible (mainly focused on the mechanical rather than intellectual processes), with varying results. Automated speech to text transcription has made significant strides in recent years, and continues to improve; this means that recorded speech, or the audio tracks of video materials containing recorded speech, can be automatically translated to text, and then methods developed for textual content can be applied (with some adjustments). Images and video have proven much more difficult -- in part because they can have meaning on so many different levels, and can concentrate a great diversity of meaning so intensely. Here intellectual analysis has been hideously labor-intensive and difficult; there are also fundamental conceptual problems about granularity and detail of description. I am reminded, for example, of the many ways and levels at which one can describe a painting of The Last Supper.

The most successful work on content-based image retrieval has, I think, occurred either in very constrained contexts (think about fingerprint matching, or face recognition) or has been limited to "vocabularies" very different from the way that most people think about images. (For example: I want images with lots of green on the bottom, blue on top, bits of yellow in the green -- this will retrieve meadows with flowers on sunny days, among other things, but it's not the way most of us usually ask for pictures of alpine meadows.)

For many kinds of nontextual materials, then, it seems that human intellectual intervention in the descriptive process is going to continue to be essential, at least for a considerable time to come. Bibliographic control of these materials is a part of this intellectual intervention to provide access. It's interesting to me that control of nontextual materials still seems to be one of the most complex and

controversial areas, perhaps in part because there is a still not fully understood confusion of objectives in the work. But this will be a critical area as we think about the context for the new millennium; here the competitors to traditional approaches -- in particular content-based retrieval -- have more limited capabilities.

I've talked about three approaches to information access that, I believe, need to be viewed as complementary rather than competing, one of which is intellectual bibliographic control. The most effective ways to use the three approaches together is still a hard research problem (albeit one that forms an essential if uncertain context for any meaningful deliberations about the future of bibliographic control). But while this synthesis develops, it is also worth exploring possibilities for shared infrastructure among the three approaches, both as a way of encouraging synthesis (and indeed even dialog among the disparate communities that may help to advance such a synthesis) and as a means of leveraging investments. I offer three areas for exploration here which should be considered as another part of the context for the new bibliographic control.

First, we know that bibliographic control is not just about rules and practices. It also depends upon a rich and complex infrastructure of authority files and classification structures. Indeed, the other approaches also use infrastructure Ð for example, lexicons, dictionaries, gazetteers and similar tools for content-oriented computational techniques, and methods to manage identity, authenticity, and reputation in the case of socially-based systems. It will be important to determine how much of this infrastructure can be shared, and leveraged, among the three approaches, and what the practitioners of each approach can do to enhance this.

Second, we must recognize the democratizing and empowering character of the networked information environment; just as anyone can become a distributor of information with a global reach, anyone can become a describer of information. Quality and trust will be as much of a problem for description of content as it is for the content itself. Metadata itself is information, and we need to be able to decide when we choose to trust it; thus many of the same tools and techniques that have become relevant to the socially based discovery of information in the digital world will also become applicable in the production and use of bibliographic metadata Ð the linkage of metadata to identities through digital signatures, the management of identities through public key infrastructure, and the manipulation of reputation related to these identities. Thus we have a specific challenge in understanding how to connect and apply the infrastructure that is being driven by the social techniques Ð and indeed by much broader developments in the networked environment, such as electronic commerce Ð to bibliographic control.

Third, I believe that as part of the massive migration of content to digital form we are approaching a crucial point in standards-setting. Digital content isn't going to be simply text (or images, or sound); rather it is going to be complex structured objects that include both the "content" -- the text, images or whatever -- and also tagged metadata associated with the content. The particular metadata elements that are available will be important both for the automation of some traditional bibliographic control functions and for the support and enhancement of content-based and social information finding systems. All of the concerned retrieval communities need to have a voice in the discussions about standards in this area (along with other interested parties, such as those concerned with rights management, and the scholars

who work with the materials). And I want to particularly highlight the linkage between these issues and issues about trust and quality -- for example, under what circumstances would bibliographic control practices countenance the automated extraction of metadata elements from a work into a bibliographic surrogate without human intellectual review and validation?

Clearly, there are opportunities for immediate and fruitful collaboration among the three communities of information finding practice, even as we strive to understand the deeper and longer-term questions about how to converge the contributions of the three communities, and how, in light of this convergence or synthesis, the practices of each individual community can be modernized, reshaped and made more effective.

We are entering a new world where content will be predominantly digital, and where it will be used, not just located, using electronic information systems. We cannot and must not attempt to map the future of bibliographic control without recognizing this. Continuing to ignore developments outside of the traditional scope of bibliographic control and to argue for business as usual -- and ever-growing funding to support business as usual -- runs the very real risk that our traditional practices may be discarded as unaffordable and of insufficient value in light of what the new technologies can offer. In my view this would be a tragedy; instead, we must concentrate on determining what bibliographic control practice can uniquely contribute, and where, when and how this contribution matters most. This means we must understand the changing context, and the economics, capabilities and limitations of the alternatives.

The economic pressures will be real as bibliographic control extends from print, where shared collaborative cataloging systems like OCLC have given us economies of scale in managing material that is acquired by many institutions, to special collections, where vast numbers of one-of-a-kind, unique items call for expensive original description. Worse, many of these items are non-textual, making them even more expensive to describe.

Finally, there is the problem of transition. Destiny may be digital, but we will be a long time reaching this destiny, and this long transitional period will call for careful management. We are already seeing print collections in our great libraries beginning to fade into invisibility for many patrons; materials available in digital form are so conveniently available, and so much more accessible through the range of retrieval systems when compared to print collections accessible only through bibliographic surrogates, and then further handicapped by document delivery considerations, that for these patrons the collection may as well only contain the digital content. While the amount of new material available in digital form is constantly growing, and there are major programs both in the noncommercial and commercial sectors to retrospectively convert print materials to digital form, this will be a slow process that will take many decades to complete. For these printed or other physical materials, bibliographic surrogates (and to some extent perhaps socially-based discovery systems) are the only means of access. What can be done to make them more visible, more accessible, to avoid partitioning knowledge into first-class (digital) information and second-class (physical) information? Bibliographic control carries a special, and heavy, burden here, and this raises serious questions about the allocation of resources for bibliographic control, and how to balance investments in bibliographic control and retrospective digitization.

This new context -- the emergence of cheap, ubiquitously available content-based retrieval approaches, and the great expansion of socially-based techniques for finding potentially relevant information -- leave us with a number of challenges in charting a future for the development of bibliographic control practices in the new millennium. What are the unique contributions of approaches based on human intellectual analysis? When is the use of intellectual analysis justified, and on what basis? What can we stop doing, or assign a lower priority to based on the assumption that content-based methods are available -- and how to our assumptions about the structure and format of the digital content that is available to these content-based retrieval systems (i.e. SGML or XML markup) shape our answers to this question?

Can we devise a spectrum of bibliographic approaches, with an accompanying spectrum of costs, to complement the content-based and socially-based approaches? Do we need to take the philosophically troublesome but perhaps pragmatic step of adopting different strategies for material that does or does not exist in digital form? How do we most effectively fuse the three approaches into information retrieval systems that are truly responsive to user needs?

The bibliographic control community cannot answer these questions alone. And they cannot shape their future without participating in a search for the answers to these questions. Redesigning bibliographic control for the new millennium will call for a new dialog among all parties and perspectives concerned with information finding that is grounded in a study of how the full array of tools and techniques now available can be applied to find information most effectively, and not in the inherent correctness or superiority of any one approach.

Library of Congress
January 30, 2001
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

LC21: A Digital Strategy for the Library of Congress

Conference program

Speakers, commentators, and papers

Conference sponsors

Conference discussion list

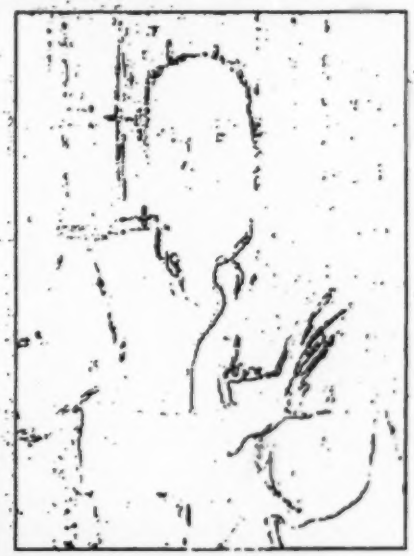
Logistical information for conference participants

Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web

sponsored by the Library of Congress Cataloging Directorate



Clifford Lynch

Executive Director,
Coalition for Networked Information

The New Context for Bibliographic Control In the New Millennium

About the presenter:

Clifford Lynch has been the Director of the Coalition for Networked Information (CNI) since July 1997. CNI, jointly sponsored by the Association of Research Libraries and Educause, includes about 200 member organizations concerned with the use of information technology and networked information to enhance scholarship and intellectual productivity. Prior to joining CNI, Lynch spent 18 years at the University of California Office of the President, the last 10 as Director of Library Automation. Lynch, who holds a Ph.D. in Computer Science from the University of California, Berkeley, is an adjunct professor at Berkeley's School of Information Management and Systems. He is a past president of the American Society for Information Science and a fellow of the American Association for the Advancement of Science and the National Information Standards Organization. Lynch currently serves on the Internet 2 Applications Council; he was a member of the National Research Council committee that recently published The Digital Dilemma: Intellectual Property in the Information Infrastructure, and now serves on the NRC's committee on Broadband Last-Mile Technology.

Full text of comments is available

Summary:

Cataloging Directorate Home Page

Library of Congress Home Page

Supporting the identification of works of interest is not the only purpose of bibliographic control, but it is certainly one of the most important and most widely relied-upon. In this paper I will consider the ways in which information finding is changing in a world of digital information and associated search systems, with particular focus on methods of locating information that are distinct from, but complementary to, established practices of bibliographic description. A full understanding of these developments is essential in re-thinking bibliographic control in the new millennium, because they fundamentally change the roles and importance of bibliographic metadata in information discovery processes.

There are three major approaches to finding information: through bibliographic surrogates, that represent an intellectual description of aspects and attributes of a work; through computational, content-based techniques that compare queries to parts of the actual works themselves; and through social processes that consider works in relationship to the user and his or her characteristics and history, to other works, and also to the behavior of other communities of users.

The first approach is familiar, and forms the basis of catalogs and abstracting and indexing, and more recently online catalogs and similar systems. The third approach is also familiar, in the form of book reviews, citation indexes, and suggestions from colleagues, but is now seeing a great creative expansion in the digital world, with its ability to create and aggregate world-wide communities of interest and to track the behavior of users. The second is fundamentally new in the digital world, where techniques based on full text searching form the basis of today's web search engines. We need to recognize that in the new millennium, for digital materials, high quality content-based computational techniques will be an inexpensive, ubiquitous, and rapidly-available default means of searching, and that powerful socially based approaches will also be widely available at little cost.

This leaves us with a number of challenges for bibliographic description in the new millennium. What are the unique contributions of approaches based on human intellectual analysis? When are they justified, and on what basis? Can we devise a spectrum of bibliographic approaches, with an accompanying spectrum of costs, to complement the content-based and socially-based approaches? How do we most effectively fuse the three approaches into information discovery systems that are truly responsive to user needs?

There is an additional set of questions that need to be considered as part of mapping the context for the new bibliographic control.

First, we know that bibliographic control is not just about rules and practices. It also depends upon a rich and complex infrastructure of authority files and classification structures. Indeed, the other approaches also use infrastructure - for example, lexicons, dictionaries, gazetteers and similar tools for content-oriented computational techniques, and methods to manage identity, authenticity, and reputation in the case of socially-

based systems. It will be important to determine how much of this infrastructure can be shared, and leveraged, among the three approaches, and what the practitioners of each approach can do to enhance this.

Second, we must recognize the democratizing and empowering character of the networked information environment; just as anyone can become a distributor of information with a global reach, anyone can become a describer of information. Metadata itself is information, and we need to be able to decide when we choose to trust it; thus many of the same tools and techniques that have become relevant to the socially based discovery of information in the digital world will also become applicable in the production and use of bibliographic metadata - the linkage of metadata to identities through digital signatures, the management of identities through public key infrastructure, and the manipulation of reputation related to these identities. Thus we have a specific challenge in understanding how to connect and apply the infrastructure that is being driven by the social techniques - and indeed by much broader developments in the networked environment, such as electronic commerce - to bibliographic control.

Library of Congress
January 31, 2001
Comments: lcweb@loc.gov